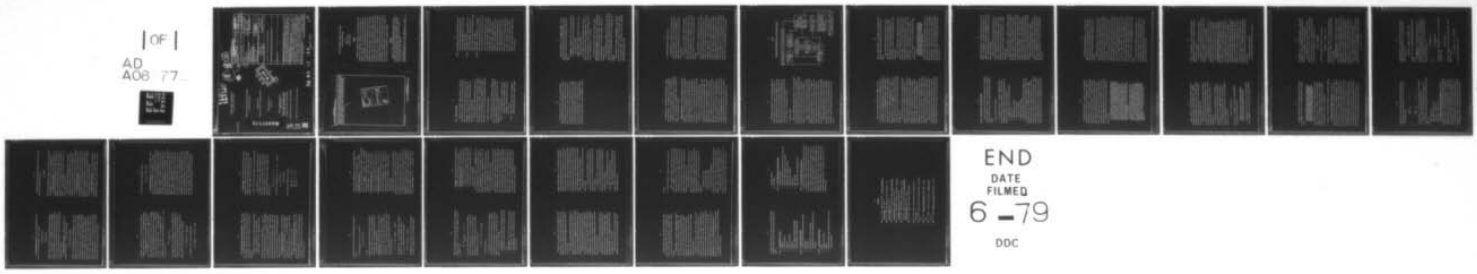
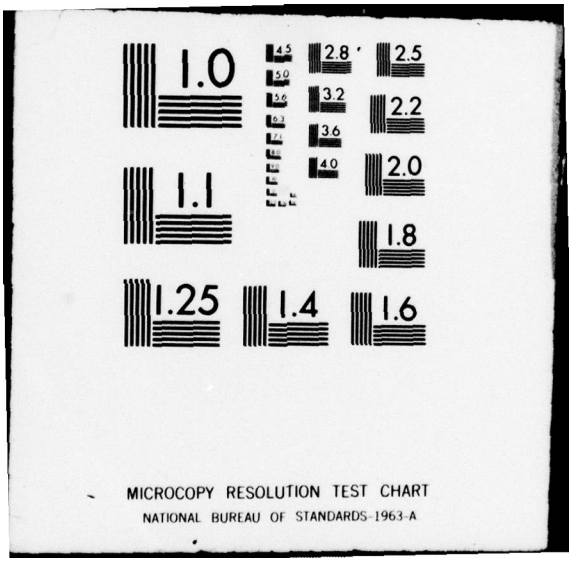
AD-A067 772 TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS F/G 12/1
STATISTICAL SCIENCE, STATISTICAL DATA MODELING, AND STATISTICAL--ETC(U)
MAR 79 E PARZEN N00014-78-C-0599
UNCLASSIFIED TR-N-6 NL



END
DATE FILMED
6-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

LEVEL II

8 9 10

TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843

INSTITUTE OF STATISTICS
Phone 713 - 845-3141

STATISTICAL SCIENCE, STATISTICAL DATA MODELING, AND STATISTICAL EDUCATION

by Emanuel Parzen

Institute of Statistics, Texas A&M University

Technical Report No. N-6

March 1979

Texas A & M Research Foundation
Project No. 3838

"Multiple Time Series Modeling and Time Series Theoretic Statistical Methods"

Sponsored by the Office of Naval Research

Professor Emanuel Parzen, Principal Investigator

Approved for public release; distribution unlimited.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: (14) TR-6 / Technical Report No. N-6
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER (9)
4. TITLE (and Subtitle): Statistical Science, Statistical Data Modeling, and Statistical Education
5. TYPE OF REPORT & PERIOD COVERED: Technical Rept.
6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): (10) Emanuel Parzen
8. CONTRACT OR GRANT NUMBER(s): (15) N00014-78-C-0599
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Texas A&M University, Institute of Statistics, College Station, TX 77843
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Code 436, Arlington, VA 22217 (11)
12. REPORT DATE: March 1979
13. NUMBER OF PAGES: 38
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): (12) 21p.
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited.
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report): NA
18. SUPPLEMENTARY NOTES: NA
19. KEY WORDS (Continue on reverse side if necessary and identify by block number): Mathematical science, mathematical modeling, professional science, statistical science, statistics, confirmatory data analysis, exploratory data analysis, statistical analysis, statistical synthesis, model identification, quantile function, density quantile function, time series analysis, statistical graduate education
20. ABSTRACT (Continue on reverse side if necessary and identify by block number):

This paper is intended to communicate to a general audience current directions in the discipline of statistics, and in my own research. It consists of three parts. Part I (Sections 1-3) explores perspectives on the future development of the science of statistics, and the importance to statisticians and society that each understand the interplay between mathematics, science, and statistics. Definitions are given of mathematical beauty and utility, soft and hard mathematics, mathematical modeling, the distinction

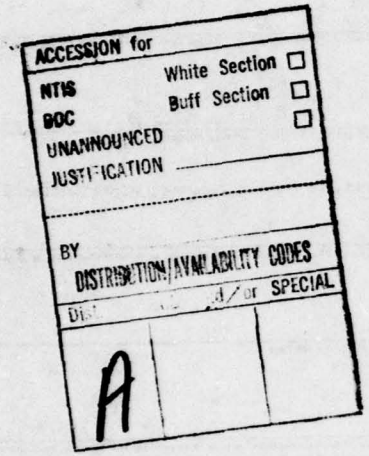
DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

between pure science and professional science, and finally statistical science. Part II (Sections 4-7) discusses the steps of statistical reasoning (including exploratory data analysis and statistical model identifications), the quantile and density-quantile approach to statistical data analysis, and time series analysis. Part III (Section 8) presents a model of statistics for organization of graduate instruction.



S N 0102- LF- 014- 6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

STATISTICAL SCIENCE, STATISTICAL DATA MODELING, AND STATISTICAL EDUCATION

by

Emanuel Parzen

Texas A&M University

Summary

This paper is intended to communicate to a general audience current directions in the discipline of statistics, and in my own research. It consists of three parts. Part I (Sections 1-3) explores perspectives on the future development of the science of statistics, and the importance to statisticians and society that each understand the interplay between mathematics, science, and statistics. Definitions are given of mathematical beauty and utility, soft and hard mathematics, mathematical modeling, the distinction between pure science and professional science, and finally statistical science. Part II (Sections 4-7) discusses the steps of statistical reasoning (including exploratory data analysis and statistical model identifications), the quantile and density-quantile approach to statistical data analysis, and time series analysis. Part III (Section 8) presents a model of statistics for organization of graduate instruction

Keywords

Mathematical science, mathematical modeling, professional science, statistical science, statistics, confirmatory data analysis, exploratory data analysis, statistical analysis, statistical synthesis, model identification, quantile function, density quantile function, time series analysis, statistical graduate education.

This research was supported in part by the Army Research Office (Grant DAAG29-78-G-0180) and the Office of Naval Research (Contract N00014-78-C-0599).

Part I. Whither or wither?

A question which I believe statisticians must face today can be succinctly phrased: "whither or wither." By "whither" I mean defining the scope and directions of the discipline most relevant to the challenges of the 1980's. By "wither" I mean the possibility of a declining role of professionals called statisticians in meeting the great demand for statistical science.

To answer the question of "whither statistics" I propose that we adopt the following definition of statistics: Statistics is a discipline at the interface of mathematics and science. Its mathematical methods are a branch of modeling mathematics; its scientific role is mainly in professional science. This part explains this definition in more detail.

1. Mathematics and mathematical modeling

In order to describe the relation of statistics to mathematics, one needs to adopt a perspective which starts with a history of mathematics and the mathematical sciences, and distinguishes five stages in the development of the mathematical sciences:

Stage 1. The rise of mathematics as an independent and purely theoretical science, and the formation of arithmetic and geometry. (Approximate dates for this stage are 500 BC - 200 AD.)

Stage 2. The completion (by about 1700) of the development of elementary mathematics (arithmetic, geometry, algebra).

Stage 3. The development (in the period 1700-1900) of the mathematics of classical physics and representing what is today called "applied mathematics"; the basic concepts are derivative and integral.

Stage 4. The development (starting in the 19th century) of contemporary pure mathematics; the basic concepts are: logic, set, function, limit, functional analysis, topology, abstract algebra and geometry.

Stage 5. The development in the 20th century of the mathematical sciences (including probability, statistics, decision theory, information theory, communication theory, systems theory, control theory, operations research, computer science, numerical analysis, mathematical programming, and optimization theory). In addition, within other disciplines (for example, management, economics, psychology, sociology, geography, education, geology, biology, engineering) there have developed mathematical subfields, usually described by names which include "mathematical" and "metrics." In the mathematical sciences, the point of view is based on using the concepts: algorithm, uncertainty or probability, optimization, data analysis, statistical inference, modeling, system, computer program, statistical design of an investigation.

The question naturally arises: is there a real philosophical distinction between statistics and applied mathematics, or is the distinction merely one of administrative convenience. The aim of this paper is to show that there is a type of statistician who reflects a real philosophical difference. The statistical scientist is not just concerned with developing applicable mathematics; he is also concerned with applying the mathematical theories of probability and statistics in a way that can be described as the "delivery of statistical care" and "developing a data-side manner." The corresponding type of mathematician is what should be meant by the phrase "mathematical scientist."

People study mathematics for its utility and beauty: utility in solving problems and beauty which provides intellectual and aesthetic satisfaction. The beauty and utility of mathematics derives from two aspects:

(1) Generality. Mathematical truths, which are facts about abstract objects rather than the real world, are very general and apply to many specific situations.

(2) Rigor. Mathematical truths are logically true and have a very high degree of conclusiveness.

As an illustration of rigor (that mathematical truth is proved by logic, not by observation) consider the Pythagorean theorem: $a^2 + b^2 = c^2$. A real triangle may appear not to exactly satisfy this property, because of errors of observing lengths a, b, c of its sides.

As an illustration of beauty, consider the proposition: Every positive integer is interesting. The proof is elegant. If the proposition is false, there exist uninteresting positive integers. Consider the smallest uninteresting integer; it is certainly interesting. There are no uninteresting integers.

The ultimate understanding of mathematics (and similarly of statistics) comes from understanding "How people do mathematics and mathematical sciences." I believe one can distinguish three levels of how, called soft, hard, and modeling.

Soft mathematics is concerned with clarification, with understanding the logical structure of a field of thought. Its main results are definitions and existence theorems. An illustration is the theorem: Every n-th degree polynomial has exactly n roots.

Hard mathematics is concerned with finding solutions, not merely proving their existence. To a mathematical concept one can distinguish three aspects: define it, compute it, interpret it. Hard mathematics is concerned with the computing and related manipulation of a mathematical concept. The interpretation of a mathematical concept is usually done in conjunction with mathematical modeling.

Mathematical Modeling is concerned with mathematical theories about mathematical concepts which can be directly interpreted as representing real phenomena in a narrow empirical area. One does this through a process called feedback between material truths (reality) and logical truth (model). An essential feature is to learn from the mistakes of the model how to build better models (one calls this "being wrong in a constructive way"). The feedback process in model building as practiced by statisticians has been well described by George Box (1976).

There seems to be great confusion and lack of understanding among the statisticians about the role and use of mathematical models. Some statisticians claim that mathematical thinking, which is concerned with finding the logically correct answer to a mathematically well-posed question, is dangerous because it could lead to the error of giving the right answer to the wrong question (this is called "an error of the third kind"). The solution is not to give up answering mathematically well posed questions based on a model, but to combine these with other techniques called exploratory mathematical analysis; exploratory analysis is concerned with framing correct questions (a question is considered adequately correct if its assumptions are adequately correct).

2. Scientific modeling and professional science

To understand the role of modeling in science one should read a recent paper "Is statistics a science?" by M.J.R. Healy (1978) which offers valuable insight into the nature of scientific modeling. I interpret Healy as distinguishing two roles of scientific reasoning: utility and beauty.

Beauty is the role of pure science: knowledge for knowledge's sake. As defined by Sir Karl Popper, science is the search for truth using constructive self-criticism (frame hypotheses and subject them to tests of ever-increasing severity). The questions of science as truth demand the answer "yes" or "no" (does or does not the theory stand up under test).

Utility is the "earthy" or "professional" role of science; it is not to discover new theory but to apply existing theoretical or empirical knowledge to the practical questions whose answers or solutions are desired by society. These practical scientists are often called workers in technology or applied science. I would like to suggest the name "professional science" since it is performed by experts practicing a profession. Many contemporary commentators (including Healy) blame the problems of lagging productivity and deteriorating quality of life in English-speaking countries on the inadequate emphasis on educating practicioners of professional science. The communication gap is not between the Two Cultures (of C.P. Snow) but among Three Cultures: the sciences (physical and life), the arts and humanities, and the professions.

Understanding the distinction (as well as achieving a golden mean) between the beauty and utility of science is the major task of all reseach supporting organizations and educational programs. Healy recommends that the education of statisticians should be guided by the fact that "If the newly qualified statistician does not instantly enter upon a teaching career, he will find that many more technologists than scientists will be anxious to take advantage of his skills."

3. Statistical science

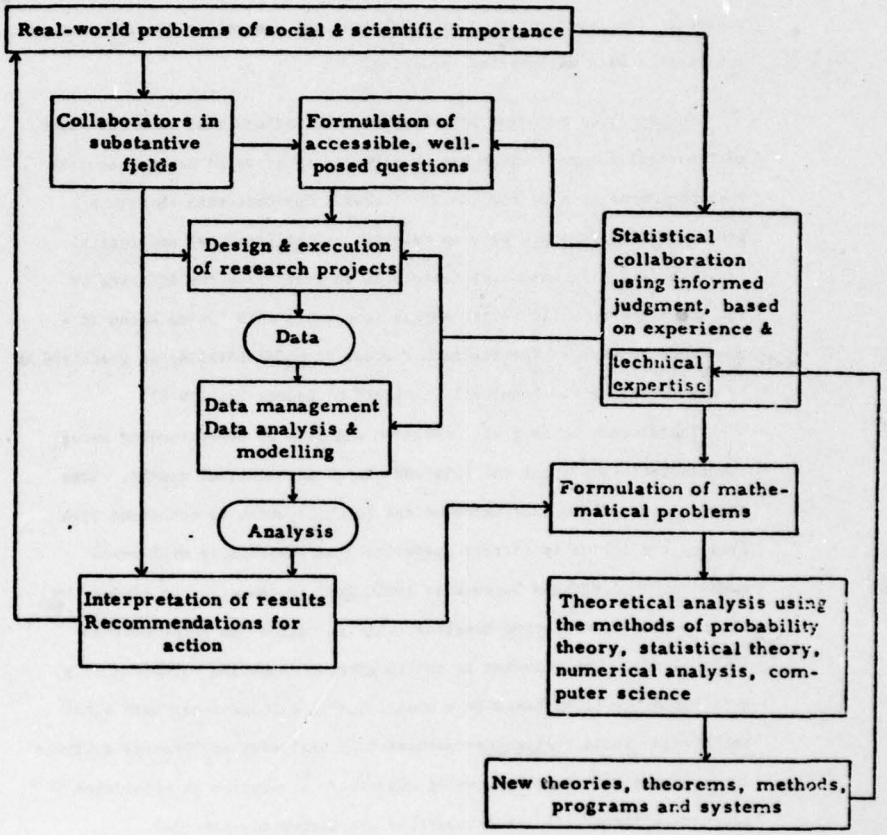
Almost all statisticians currently writing on the future of statistics [Box (1976), Healy (1978)] emphasize that we need to train statisticians who are able to conduct collaborative research with professional scientists. The appropriate name for such statisticians seems to be "statistical scientist." Their roles can be depicted by a flow chart (Figure A).

Research in statistics can be pure (concerned with the development of new methods) or professional (concerned with the application of known methods to solving problems important to society). To be qualified to be a statistical teacher or consultant in the decade of the 1980's, one requires a broad education in many fields of statistics (beyond the areas of experimental design and linear models which in the past often was the major part of the total education of many applied statisticians).

All scholars and professionals need to use statistics creatively, and therefore need to learn it in ways that stimulate an appreciation of its roles. It seems to me clear that every educated person should learn about statistical thinking; the proportion who do so now is far too low.

Figure A

Flow Chart Depicting the Roles of a Statistical Scientist



A way to provide insight into statistical thinking is to point out that the basic statistical question in any investigation of data is to distinguish between fluctuation (changes in values observed explained by their variability) and non-stationarity (changes in values observed due to changes in statistical parameters). When non-stationarity is observed, one has to determine if it is real or spurious; real if it is due to a change in the nature of reality; spurious if it is due to a change in our process of measuring reality. Thus a significant increase in crime statistics may be real (there is more crime) or spurious (the system for reporting crimes has improved).

What distinguishes the statistician from other professionals employing statistical methods is his commitment to the consistency of his methodology; consequently, he must examine any procedure for data analysis to see if it abides by the principles he would use in other circumstances seemingly unrelated to the problem at hand. The goal of a consistent methodology is mainly achieved by translating real problems to problems of probability and statistics; the latter are stated mathematically (axiomatically) using concepts that are given meanings of their own that do not depend on particular applications. Axiomatically developed concepts provide formal analogies between applications which are themselves totally different but which in certain theoretical aspects can be treated similarly.

Because the concepts of statistical data analysis possess meanings of their own which do not depend on particular applications, statistical theory provides mathematical procedures which can be used to solve problems arising in diverse fields. Thus, statistics is an inter-disciplinary field

since it provides a medium for the exchange of ideas between statisticians and research workers concerned with quite different subject matters.

A basic question which statisticians have failed to resolve is what should be their role in an information analyzing world where the availability of statistical computing program libraries could lead investigators to believe they could, and indeed should be, their own statisticians.

The answer to this question in my judgement is for statisticians to stress the insufficiency of model analysis without model checking and model synthesis. The latter cannot be routinized, and requires a broad understanding of statistical theory and methods. I would like to emphasize that I strongly believe that statistical computing, using quality software, is an essential part of statistical education.

In a 1967 conference on the future of statistics (Watts (1968)), Watts writes (p. xi):

> A measure of the effectiveness of our communication with others was given by John Tukey who contended that "If statistics is to be effective, its techniques must be kept easy enough to use, the test of 'easy enough' being whether or not they ARE used."

I believe that statistical computing has changed our criterion for what makes a statistical technique effective; what makes a technique easy is not how easy it is to compute by hand, but how easy it is to communicate its interpretation. I believe the latter aim is best achieved by comparisons of graphs of smooth curves (representing our parameter estimators and our models) with wiggly curves (representing graphical summaries of the observed data).

Part II. Some Frontiers of Statistical Science

This part outlines some recent advances in statistics that illustrate what are the advantages to a scientist of collaborating with a professional statistician broadly trained in both standard and nonstandard statistical methods.

4. Stages of statistical data analysis

The discipline of statistics is currently experiencing a healthy agitation concerning its theories and methods. I believe that one can distinguish six stages of statistical data analysis which differ in the nature of the assumptions required; these stages are:

1. Confirmatory data analysis and statistical inference,
2. Goodness of fit,
3. Robust statistical inference,
4. Exploratory data analysis,
5. Statistical data modeling,
6. Adaptive statistical inference.

Confirmatory data analysis is a modern name for the important field of statistical reasoning (traditionally called "parameter estimation and hypothesis testing") that is concerned with estimating and testing hypotheses about the values of parameters (denoted θ) of a postulated model for the true unknown probability law of the observed data. A problem of confirmatory data analysis starts with a family of possible probability laws for observations X denoted (P_θ, $\theta \in$ parameter-space). The family of possible probability laws is based on prior knowledge (usually theoretical

considerations which restrict the possible probability laws to a certain family) or prior experience. The statistical questions answered by confirmatory data analysis are considered to be <u>well posed</u> questions because all of the possible probability laws for the data are defined by the model, which is called an ideal model.

Goodness of fit tests arise when one postulates an ideal model for the data and one would like to test the hypothesis that the model fits the data. In confirmatory statistical inference, to optimally test an hypothesis one needs to specify an alternative hypothesis. In goodness of fit theory, one does not postulate a model alternative to the ideal model being tested.

Robust statistical inference is concerned with estimating parameters by methods which are not sensitive to slight deviations in the ideal model (for example: the data is assumed to be normal but in fact may have a distribution symmetric about its median with longer tails than a normal, or the data has some outlying values). Robust inference theory criticizes the mean and variance as estimators of location and scale, and least squares as an estimation principle for linear regression coefficients, and suggests alternatives. However, the decision as to which alternative estimators to use in practice requires one to learn from the data a model to be assumed for the probability distributions of the data.

One can discern another guiding principle of robust statistical inference: fit a parametric model to the data which the model fits well, and identify the remaining data as having non-standard characteristics. Often the most interesting part of the analysis is the identification of the "bad" data.

"Exploratory data analysis" is the name of a very interesting book by John Tukey (1977) and also the name of an approach to statistical reasoning which Tukey has pioneered. It seems to be an approach to data analysis which avoids statistical modeling. A discussion of attitudes towards the nature of exploratory data analysis is given in the discussion of my paper on non-parametric statistical data modeling (Parzen (1979)). There can be no doubt that John Tukey has been the prophet of the development of modern statistical data analysis. An article in the popular magazine Fortune (published in February 1964, fifteen years ago!) by George Boehm and entitled "The Science of Being Almost Certain" concludes as follows:

> The opportunities for the brilliant statistical analyst are likely to increase sharply within the next few years, according to John Tukey, who is a mathematics professor at Princeton and a member of the staff of Bell Telephone Laboratories. He foresees the development of new methods of statistical analysis that will make more use of freewheeling human judgment and intuition. Tukey wants to see data analysis turned into more of a creative art, in which the statistician can "listen to what the data is trying to tell him." He urges his fellow statisticians not to plan experiments too rigidly but to keep an open mind and let preliminary results feed back into the analysis. They should, in his opinion, appraoch the data with some definite questions in mind, but if after some study they see something suggestive, they should expand and revise their list of questions.
>
> This way of doing business will require intimate cooperation between the human mind and the computer. The first stage will be to have the computer break down data in a number of ways and perhaps display summaries graphically on a screen so that the statistician can get a feel for the problem. In the next few decades even this personal feeling may be automated to some extent as far more sophisticated computer programs are devised. Then someday perhaps the computer will automatically try to evaluate hundreds of different statistical approaches and present to the statistician only the most interesting ones. At that stage, statistics may become truly an art, calling for the utmost in the exercise of human imagination.

As exploratory data analysis matures, it is to be hoped that it becomes not an art but a science in which methods are explained not only by how they work, but also by why they work. Criteria need to be developed by which proposed methods may be evaluated, and either grow or die.

Statistical data modeling is a field of statistical reasoning that assumes that a model for the data is not available from prior theory; then one seeks to learn the model by a process called statistical model identification (whose relationship to exploratory data analysis remains to be clarified). A problem of statistical model identification (or statistical synthesis) starts with data for which little or no prior experience is available (or for which one may want to disregard the available prior knowledge). One seeks to fit a model (family of probability laws) to the data. Statistical model identification seeks a statistical solution to the problem of parametrization (or model specification). It is interesting to compare statistical design and statistical model identification. Statistical design seeks to learn by observing a phenomenon under changing conditions by seeing what happens when something is varied in a controlled way. Statistical model identification starts with a body of collected data and seeks to learn by analyzing the data under changing models.

My recent research on statistical data modeling (described in the next two sections) attempts to include traditional non-parametric statistical inference as a special case. It provides new approaches to non-parametric probability density estimation, but emphasizes estimation of quantile and density-quantile functions.

Adaptive statistical inference attempts to combine confirmatory data analysis with statistical model identification; it estimates parameters using estimators which are adaptive in the sense that their coefficients (which depend qualitatively on the model being assumed) are learned or estimated from the data.

The significant feature that I believe will make the discipline of statistics very different in the next decade from what it has been in the past is our increasing ability to check the validity of assumptions about models for observed data that we formerly accepted unquestioningly.

The known family P of possible probability laws assumed by a statistical model is called a parametric family if one assumes that there is a space, called the parameter space (which in general is multi-dimensional) which can be used to label P in the sense that each point θ in parameter-space corresponds to exactly one probability law P_θ in P; we then write

$$P = \{P_\theta,\ \theta \in \text{parameter-space}\}$$

and say that θ is a parameter representing the state of nature or true probability law of the observations.

Parameters may represent a classification introduced by the statistician or may be imposed by prior theory.

The parameter space for a given family P of possible probability laws is not unique; one can define a multitude of parameter spaces for

a given P. In my judgement, it is useful to philosophically divide parameter spaces into two categories: (1) synthetic or statistical and (2) analytic or structural.

If we are interested only in identifying the true probability law as a member of P, and the parameter space is chosen as one possible representation of the members of P, we call the parameter space statistical (or synthetic) and call θ a statistical parameter.

If the parameters θ have a physical interpretation related to the random mechanism generating the observations, we call the parameter space structural (or analytic) and call θ a structural parameter.

Statistical theory and methods are said to be concerned with confirmatory data analysis and behavior in the face of uncertainty when a model for the uncertainty can be assumed known. This field of statistics covers both its optimal decision-making applications and its optimal information extraction applications. However, in many scientific applications there is yet a third role of statistics: learning from data relationships (regularities, patterns, even peculiarities) contained therein without assuming a precise model, although assuming that the relationships are not deterministic but need to be expressed in terms of the concepts of probability theory. This field of statistics is treated by "statistical model identification" and "exploratory data analysis." The information extraction applications of statistics are thus divided into (at least) two classes, to which I also give the names:

(1) statistical analysis (confirmatory)

(2) statistical synthesis (identification or exploratory).

An important phase of statistical reasoning is statistical design of investigations; what the latter encompasses is described by Federer (1978):

> Description of variables and populations, measurements and measuring instruments, treatment design, experiment design, sequential design, sample survey design, model building design, determination of sample size, and principles and properties of statistical designs. These items must be considered for either planned or unplanned investigations.

As the flow chart in Figure A indicates, a statistical research project consists of three phases: (1) design of the experiment and collection of the data, (2) data management, statistical analysis, and probability modeling, and (3) interpretation of the results leading to recommendations for decisions or controls to be implemented.

5. The quantile and density-quantile approach to statistical data modeling

(I) Probability based statistical data analysis. When asked the exploratory question "here is a data set; what can be concluded," what we desire to draw conclusions about is the probability distribution from which the sample purports to be a representative sample. Standard statistical theory is concerned with inferring from a sample the properties of a random variable that are expressed by its distribution function $F(x) = Pr[X \leq x]$ and its density function $f(x) = F'(x)$, which are functions on $-\infty < x < \infty$. The quantile and density-quantile approach to statistical data modeling [Parzen (1979)] is based on the observation that, both for theory and practice, the appropriate way to compute and graphically present a probability distribution function is to compute and plot the quantile function $Q(u) = F^{-1}(u)$ and the density-quantile function $fQ(u) = f(Q(u))$, which

are functions on $0 \leq u \leq 1$. The qualitative behavior of Q and fQ leads naturally to a division of probability laws into types based on their tail behavior, their symmetry, and their differentiability.

It is a fundamental fact that a random variable X with quantile function $Q(u) = F^{-1}(u)$, where $F(x) = P(X \leq x)$, is identically distributed as Q(U), where U is uniform on (0, 1). Consequently, expectations of any function g(X) can be expressed as an integral on $0 \leq u \leq 1$:

$$E[g(X)] = \int_0^1 g(Q(u))du \ .$$

In particular the mean μ and variance σ^2 of X are given by

$$\mu = \int_0^1 Q(u)du \ , \quad \sigma^2 = \int_0^1 (Q(u) - \mu)^2 \, du \ .$$

These formulas provide definitions of μ and σ^2 valid for both continuous and discrete random variables. However, the discussion henceforth assumes that X is continuous.

(II) Descriptive statistics. Averages such as the mean or median (or even the trimmed mean) do not tell us all we need to know about the data we are analyzing. We need to know the probability distribution of the data. Greatest insight into the probability distribution generating a sample is provided by computing and graphing its sample quantile (or percentile) function $\tilde{Q}(p)$, $0 \leq p \leq 1$; $\tilde{Q}(p)$ is a number such that 100p% of the data is below $\tilde{Q}(p)$ and 100(1 - p)% of the data is above $\tilde{Q}(p)$. The descriptive statistics of a sample can be displayed by a Quantile-Box plot.

When a sample of size n is available as a set of numbers (rather than a histogram) one should form the order statistics of the sample, denoted

$$X_{1;n} \leq X_{2;n} \leq \cdots \leq X_{n;n}$$

We define $\tilde{Q}(u)$ to be a piecewise linear function connecting the values defined at points which are multiples of $1/(n+1)$ or odd multiples of $1/(2n)$; we define either

$$\tilde{Q}(\tfrac{j-\frac{1}{2}}{n}) = X_{j;n} , \qquad j = 1, 2, \ldots, n ,$$

or

$$\tilde{Q}(\tfrac{j}{n+1}) = X_{j;n} , \qquad j = 1, 2, \ldots, n .$$

The value of $\tilde{Q}(u)$ at $u = 0$ and 1 needs to be specified; we usually define it to equal the sample minimum and maximum respectively.

We argue that all the statistical information in a sample is contained in (and indeed is easily extracted from) the sample quantile functions. Examples are

Sample median $\tilde{Q}(0.5)$,

Sample quantiles $\tilde{Q}(0.25)$, $\tilde{Q}(0.75)$,

Sample mean $\bar{X} = \int_0^1 \tilde{Q}(u)du$,

Sample variance $\int_0^1 (\tilde{Q}(u) - \bar{X})^2 \, du$.

(III) Estimation of location and scale parameters. The model for a distribution function $F(x) = F_0(\frac{x-\mu}{\sigma})$ where $F_0(x)$ is a specified distribution function and μ and σ are location and scale parameters to be estimated, is equivalent to a model for the quantile function $Q(u) = \mu + \sigma Q_0(u)$, where $Q_0(u)$ is the quantile function corresponding to $F_0(x)$. The parameter estimation approach to statistical inference seeks efficient estimators $\hat{\mu}$ and $\hat{\sigma}$ of μ and σ respectively; one can then form an efficient estimator $\hat{Q}(u)$ of $Q(u)$ by $\hat{Q}(u) = \hat{\mu} + \hat{\sigma}Q_0(u)$. A computationally tractable way to compute efficient estimators $\hat{\mu}$ and $\hat{\sigma}$ in the regular case where these estimators are

asymptotically normal, is by regression analysis of the "continuous parameter" time series $f_0Q_0(u)\tilde{Q}(u)$, $0 \le u \le 1$, which obeys the model (in the "regular" case)

$$f_0Q_0(u)\,\tilde{Q}(u) = \mu\, f_0Q_0(u) + \sigma\, Q_0(u) f_0Q_0(u) + \sigma\, B(u)$$

where $B(u)$, $0 \le u \le 1$ is a Brownian Bridge (or Pinned Brownian Motion) process.

(IV) Goodness of fit. To test the hypothesis H_0 that the true quantile function $Q(u)$ is of the form $Q(u) = \mu + \sigma Q_0(u)$, where Q_0 is specified, one tests the hypothesis that a suitably defined density $d(u)$, $0 \le u \le 1$, satisfies $d(u) \equiv 1$. We call $d(u)$ a preflattened quantile-density function, and define it by

$$d(u) = \frac{1}{\sigma_0} f_0Q_0(u)\, Q'(u)$$

where

$$\sigma_0 = \int_0^1 f_0Q_0(u)Q'(u)du$$

is a measure of scale. When $f_0Q_0(u)Q(u) = 0$ at $u = 0$ and 1, one has by integration by parts

$$\sigma_0 = \int_0^1 J_0(u)\, Q(u)du ,$$

where $J_0(u) = -(f_0Q_0)'(u)$ is called the score function of Q_0. Another formula for $J_0(u)$ is

$$J_0(u) = \psi(Q_0(u)) , \quad \psi(x) = -\frac{d}{dx} \log f_0(x) .$$

One calls $\psi(x)$ the Fisher score function.

(V) Robust estimator of scale. It seems to me remarkable that there seems to be a universal estimator of scale σ, given by

$$\tilde{\sigma}_0 = \int_0^1 f_0 Q_0(u)\, \tilde{Q}'(u)du$$

$$= \int_0^1 J_0(u)\, \tilde{Q}(u)du$$

I believe that one can unify the diverse approaches to robust estimation called L, M, and R estimation: the estimator is chosen by an optimization criterion as one minimizing a suitable measure $\tilde{\sigma}_0$ of deviation of residuals from zero. Robust statistical inference can be developed from two points of view which can be called maximum robust likelihood estimation and minimum robust deviation estimation.

(VI) Robust estimation of location. In general an average is an estimator μ^* of location μ. An average can be represented

$$\mu^* = \int_0^1 W(u)\, \tilde{Q}(u)du = \sum_{j=1}^{n} W_j X_{j;n}$$

where $W(u)$ is a suitable weight function (integrates to 1), and W_j is the average of $W(U)$ over an interval [such as $(2j - 1)/2n \leq u \leq (2j + 1)/2n$ or $j/(n + 1) \leq u \leq (j + 1)/(n + 1)$]. I describe this formula as leading to a definition of statistics: "statistics is arithmetic done by ranking before adding." Confirmatory statistical inference chooses the weight function $W(u)$ by assuming an ideal model. Then $W(u) = f_0 Q_0(u) J_0'(u)$, normalized to integrate to 1. One can define weight functions $w(x)$ such that $W(u) = w(Q_0(u))$. This observation can be used to motivate the construction of robust estimators of μ when one assumes a semi-ideal location and scale parameter model $Q(u) = \mu + \sigma\, Q_0(u)$ where $Q_0(u)$ is an unspecified quantile function assumed to be symmetric about its median.

The weight function $w(x)$ for robust estimation of location is often chosen to be a robust representative of

$$w(x) = \frac{\psi(x)}{x},$$

where $\psi(x)$ is the Fisher score function. A robust window (weight function) that I recommend is: $w(x) = (1 + \frac{1}{m} x^2)^{-1}$ for a suitable value of m.

(VII) Exploratory data analysis. We cannot safely assume that the distribution of the variable X represented by the sample is normal or even symmetric. Using the Quantile-Box plot we can visually check normality, and diagnose to which one of several qualitative types the distribution of X belongs: (i) unimodal and symmetric, (ii) unimodal and skewed, (iii) bimodal or multimodal, (iv) discrete.

When data is reported as a histogram more insight is obtained by plotting not the histogram, but the histogram-quantile function (defined as a function on $0 \leq u \leq 1$ whose values at u is the histogram value at $x = \tilde{Q}(u)$, where one defines $\tilde{Q}$ to be the inverse of the cumulative histogram).

(VIII) Statistical data modelling. The quantitative estimation of fQ can be accomplished non-parametrically by a variety of smoothing or density estimation methods. There is a procedure which can be especially recommended, called autoregressive estimation of a prefalttened quantile-density function $d(u) = f_0 Q_0(u) q(u) \div \sigma_0$, where $f_0 Q_0(u)$ is suggested by a goodness of fit hypothesis H_0 or by exploratory data analysis.

(IX) Adaptive statistical inference. The quantitative estimation of the score function $J(u) = -(fQ)'(u)$ can be accomplished by the method of autoregressive estimation in a way that provides an analytic formula for

the estimator. Estimators of $J(u)$ can be substituted in confirmatory statistical inference formulas for $\hat{\mu}$ and $\hat{\sigma}$ to provide adaptive estimators.

(X) The concentration problem, which arises in many fields of social science, is an example of a problem clarified by thinking in terms of quantile functions and density-quantile functions. There is a non-negative variable X representing measurements such as acreage of a farm, population of a city, number of publications of an author, and so on. One observes n units (farms, cities, or authors), and $X_1, \ldots, X_n$ are their measurements. Let $S = X_1 + \ldots + X_n$ be the total. Let the measurements be ranked (arranged in increasing order) with order statistics $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$. Form for p in $0 < p < 1$

$$\tilde{L}(p) = \sum_{j \leq p(n+1)} X_{(j)} \div S .$$

One interprest $\tilde{L}(p)$ to be the fraction of the total S contributed by the smallest 100p% of the units being measured. The accepted measure of concentration is a functional, such as the integral, of the curve $\tilde{L}(p) - p$, $0 \leq p \leq 1$; one calls $\tilde{L}(p)$ the Lorenz curve. The Lorenz curve $L(p)$ of a random variable X with quantile function $Q(u)$ is defined by

$$L(p) = \frac{1}{\mu} \int_0^p Q(u)du , \quad \mu = \int_0^1 Q(u)du .$$

It is related to a test statistic for exponentially

$$D(p) = \int_0^p d(u)du , \quad d(u) = \frac{1}{\mu} (1 - u)q(u)$$

by the identity (proved by integration by parts)

$$D(p) = L(p) + \frac{1}{\mu} (1 - p)Q(p)$$

(XI) Continuous regression. One of the basic problems of statistics is to study the relationship of a variable Y to a variable X. From a practical point ov view, it is often meaningless to speak of E[Y], the mean of Y. The mean of Y is most meaningful when it is regarded as depending on the value of X; we denote it by $E[Y|X]$, called the conditional mean of Y given the value of X. By determining relationship of Y to X one might mean determining $E[Y|X]$. More generally, one should determine the conditional quantile function of Y given X, denoted $Q_{Y|X}(u)$. For ease of analysis, the statistician often quickly postulates that $E[Y|X]$, and $Q_{Y|X}(u)$, are linear functions of X. One has always been advised to avoid the tendency to compute linear relationships without ever testing if the real relationship is non-linear. Our ability to act on this advice in practice is only now becoming available, using non-parametric regression analysis. Continuous regression is the name given to this analysis when the data arises as a scatter plot (X_i, Y_i) of values, and for each X value there are only one or a few corresponding Y values. When many Y values are available for a given X value, we call the problem discrete regression or the k sample problem.

(XII) Discrete regression and k-sample problems. Consider a numerical (continuous) measurement Y made on units which can be classified into discrete groups. Denote the group to which a unit belongs by X; it takes only k possible values (which often are numerical, or at least ranked). Examples are:

Y = salary of a professional, X = number of years in the profession;

Y = investment in farm machinery; X = age of the farmer;

Y = potato consumption of a man; X = income class of the man.

For each value j of X, one has n_j measurements $Y_{1,j}, \ldots, Y_{n_j,j}$ of Y values of units having X value equal to j. For each j, one could compute $\bar{Y}_{\cdot j}$ (the sample mean of the Y values corresponding to X = j) and more generally the sample quantile function $\hat{Q}_j(u)$. Classical statistics computed and compared (and hopefully graphed) the means $\bar{Y}_{\cdot j}$. Modern data analysis would plot and compare the quantile-box plots of $\hat{Q}_j(u)$. One would seek to determine graphically if the conditional means, medians, quartiles of Y given X are linear in X.

(XIII) Comparison distribution functions. To test the equality of several distributions (corresponding say to the conditional quantile functions $Q_j(u)$ of Y given X = 1, ..., k) one would define suitable comparison distribution functions D(u), $0 \leq u \leq 1$; the basic themes of this approach are described in the next section.

(XIV) Tests for independence and multivariate modeling. The theory may be extended in several ways to such problems. Alternatives to the hypothesis of independence of X and Y can be expressed as hypotheses for the comparison of the conditional distribution of Y given X with the unconditional distribution of Y.

(XV) Bayes theorem. Conceiving of quantile functions as the basic expressions of a probability distribution may also be a useful concept for Bayesian statisticians. Bayes theorem appears very elegant when expressed in terms of quantile functions [see Parzen (1979)].

6. Density estimation as universal approach to statistical data analysis

A major theme of our work is: statistical hypotheses can often be transformed to hypotheses about a suitably defined distribution function D(u), $0 \leq u \leq 1$ on the unit interval, in such a way that a null hypothesis transforms to D(u) = u. Then, in addition to the distribution function domain, one has available the domain of the density function

$$d(u) = D'(u) , \quad 0 \leq u \leq 1 ,$$

and the Fourier transformation $\phi(v)$, $v = 0, \pm 1, \pm 2, \ldots$ defined by

$$\phi(v) = \int_0^1 e^{2\pi i u v} \, dD(u) ,$$

$$= \int_0^1 e^{2\pi i u v} \, d(u)du .$$

The following null hypotheses are equivalent:

$$H_0(D): D(u) = u , \quad 0 \leq u \leq 1$$

$$H_0(d): d(u) = 1 , \quad 0 \leq u \leq 1$$

$$H_0(\phi): \phi(v) = 0 , \quad v \neq 0$$

Statistical data analysis using the foregoing approach proceeds as follows:

1. Form raw estimators $\tilde{D}(u)$, $0 \le u \le 1$ of $D(u)$ and

$$\tilde{\phi}(v) = \int_0^1 e^{2\pi iuv}\, d\tilde{D}(u)$$

of $\phi(v)$.

2. Determine the asymptotic distribution theory of the stochastic process $\tilde{D}(u) - D(u)$, $0 \le u \le 1$. It often occurs that under the null hypothesis $D(u) = u$, the asymptotic distribution of $\tilde{D}(u) - u$ is a Brownian Bridge stochastic process.

3. Parametric models. When parametric models are available for the distribution functions of the data, there is a corresponding parametric model $D_\theta(u)$ for the true $D(u)$ function. Suppose θ has k components $\theta_1, \ldots, \theta_k$; one may be able to show that for values θ_j close to 0 (representing the null hypothesis) one has an approximate linear representation

$$D(u) = u + \theta_1\Delta_1(u) + \theta_2\Delta_2(u) + \ldots + \theta_k\Delta_k(u)$$

Then estimators $\hat{\theta}_1, \ldots, \hat{\theta}_k$ can be obtained from a (continuous parameter time series theoretic) regression of $\tilde{D}(u) - u$ on $\Delta_1, \ldots, \Delta_k$. A smooth $D(u)$ function is formed by

$$\hat{D}(u) = u + \hat{\theta}_1\Delta_1(u) + \ldots + \hat{\theta}_k\Delta_k(u)$$

By comparing the fit of $\hat{D}(u)$ to $\tilde{D}(u)$ one can obtain goodness of fit tests for parametric models. One obtains most powerful

tests of the null hypothesis $D(u) = u$ by using chi-square distributed statistics which are quadratic forms in $\hat{\theta}_1, \ldots, \hat{\theta}_k$.

4. Non-parametric tests based on D. To test the null hypothesis $D(u) = u$, one can use test statistics such as $\tilde{D}(0.5) - 0.5$, $\int_0^1 J(u)d\tilde{D}(u)$, $\int_0^1 (\tilde{D}(u) - u)du$, $\max_{0<u<1} |\tilde{D}(u) - u|$, $\int_0^1 (\tilde{D}(u) - u)^2du$. The asymptotic distributions (under the null hypothesis) of these test statistics are available since $\tilde{D}(u) - u$ is then a Brownian Bridge process.

5. Non-parametric tests based on $|\tilde{\phi}|^2$. To test the null hypothesis $\phi(v) = 0$ for $v \neq 0$, one can use the plot of $|\tilde{\phi}(v)|^2$, or test statistics of the form $\sum_{v=1}^{m} k(v)|\tilde{\phi}(v)|^2$ for suitable choices of $k(v)$.

6. Non-parametric estimation of the density d. One would like to not only test the null hypothesis but to estimate $D(u)$ by a smooth function $\hat{D}(u)$ which "parsimoniously" fits $\tilde{D}(u)$. This goal can be accomplished by various approaches to forming an estimator $\hat{d}(u)$ of the density $d(u)$; we recommend the autoregressive spectral approximation approach.

The simplest example of a $D(u)$ function arises in the problem of testing the equality of two distribution functions F and G. Then one could choose $D(u) = F(G^{-1}(u))$ with raw estimator $\tilde{D}(u) = \tilde{F}(\tilde{G}^{-1}(u))$ where $\tilde{F}$ and $\tilde{G}$ denote sample distribution functions. One may be able to introduce $\tilde{D}(u)$ even when there is no obvious definition of $D(u)$. Consider samples from two populations which do not consist of measurements (numbers) but nevertheless can be ranked. Assign to an observation its rank when the two samples are combined. Let N be the number of items in the two samples combined. For $i = 1, \ldots, N$, define $Z_i = 1$ if the

member of the combined sample with rank i comes from the first population; otherwise, define $Z_i = 0$. Let m be the number of members of the first population. Define

$$\tilde{D}(u) = \frac{1}{m} \sum_{i \le u(N+1)} Z_i , \quad 0 \le u \le 1 ;$$

it is a purely discontinuous distribution function with jumps of size 1/m at the points $u = i/(N + 1)$, $i = 1, \ldots, N$, where i is such that $Z_i = 1$. By comparing $\tilde{D}(u) - u$ with 0 one can test the null hypothesis that the two populations are equal; a test statistic corresponding to the Wilcoxon test is

$$T = \int_0^1 \{\tilde{D}(u) - u\}du = 0.5 - \int_0^1 u \, d\tilde{D}(u) .$$

7. Time series analysis

No discussion of current trends in statistics would be complete without emphasizing the importance in the education of a statistician of the study of time series analysis. To understand what are the aims and means of time series analysis, one should begin with a study of some of its classic practical problems:

1. Trend estimation. Given a wiggly curve representing observations over time, pass a smooth curve through it representing long term trends.
2. Cycle estimation. Given a time series of deviations from trend, pass a smooth curve through it representing cyclic components.
3. Spectral estimation. Given a stationary time series, estimate its spectral density.

4. Forecasting. Identify adequate formulas for forecasting future values of a time series from values up to a given time.
5. Relations between two time series. The relations between two time series are usually dynamic, in the sense that one desires to relate the value of Y(t) to values of X(s) at a set of times s. One seeks a model which identifies the lags (times at which X is most related to Y) as well as the coefficients and functional form of the relationship.

Time series analysis is a blend of: (1) probability theory in its study of stochastic models of dynamic processes evolving in time and space, (2) statistical theory in its study of methods of inference for analyzing and synthesizing models, (3) numerical analysis in its study of methods of handling large masses of data and efficient computer algorithms, and (4) systems theory in its study of the representation of time series as inputs and outputs of systems.

Time series analysis is of interest to researchers in the physical sciences (geophysics, oceanography, atmospheric physics), engineering sciences (communication theory, control theory, aerospace studies, structural vibration studies, acoustics), biological sciences (biorhythms), medicine (EEG and EKG analysis), social sciences, economics and management science (business cycles, econometrics, forecasting). We are in a "golden age" of time series analysis; its potential applications are to problems of great public concern, its methods are being systematically and routinely applied, and its instrumentation (hardward) and computer programs (software) are important industries.

There seems to be no limit to the fields of applications of time series analysis. In fields whose theoretical development is not too advanced, empirical time series analysis can play a useful exploratory role, describing measurements and suggesting models and relations to be fitted to them.

In the physical sciences (especially those whose measurements are made in the field rather than the laboratory) empirical time series analysis is mainly used for estimation of parameters of models available from physical theory (among the sciences which extensively apply time series analysis are those studying the earth and the solar system).

Time series analysis and statistical signal processing have as their starting point data which are to be "processed" or "analyzed" or "modelled." The data are time series (observations associated with successive points of time). The data can be regarded as a sample path or trajectory or realization of a stochastic process (a probability model which describes probability distributions of all the time series that could have been observed).

Time series analysis analyzes and synthesizes stochastic models to describe and control the mechanisms generating a time series and relating various time series. It also treats problems of system analysis, system identification (modelling), and system optimization for stochastic systems (that is, systems whose inputs and/or outputs are stochastic signals or time series).

Time series analysis fits models to data fields and to systems, employing not only statistical and probability methods, but also

methods of numerical analysis. It is concerned with the problem of fitting or approximating experimental data in terms of polynomials, harmonics and other families of fitting functions. It aims to develop efficient computer programs in which choices of method are based, as far as possible, on statistical attitudes concerning the interpretation of the answers and in which insight into the answers obtained is provided by comparing them with the answers one obtains to other questions one might have asked. Recursive estimation procedures, and the interplay of time average and ensemble average representations and approximations, play important roles in time series analysis.

Among theories of data analysis, time series analysis has several special features:

1. The data is usually non-experimental in the sense that it is nature rather than the observer who designs the experiment which is being observed; often, the observer must passively accept the time history he observes.

2. The observer often has no pre-conceived model for the data (so that he hardly has classical estimation and testing problems) but rather seeks to obtain clues from the data as to what models should be considered for investigation.

3. Criteria for choosing between models can be formulated in terms of ideas of prediction and spectral fitting.

4. Indeed, in many problems one is not mainly concerned with finding a model for the mechanisms generating the series but only with forecasting the series.

The basic methods of time series analysis provide ways of thinking for solving problems which do not arise as the modeling of data fields or systems. They are being applied to nonparametric methods of statistical data analysis, and to numerical solutions of integral equations.

An understanding of the concepts of autocorrelation and spectrum today needs to be part of the education of every researcher. The fields in which these concepts are currently being applied include: art, history, political science, psychology, sociology, economics, management, finance, biology, medicine, geophysics, meterology, oceanography, radar, sonar, space travel, computer science, pattern recognition and image reconstruction science, analysis and synthesis of speech and written text.

References to my own research in time series analysis can be found in my recent paper "Time Series Modelling, Spectral Analysis, and Forecasting." This paper presents a strategy for building models for an observed time series which fulfill three aims: to provide time domain models which can be interpreted in terms of trend and seasonal components, provide forecasts, and provide spectral estimators. Our time series modeling strategy attempts to achieve these aims by using the concepts of predictability and fitting the spectral distribution function. The approach described could be called: "the autoregressive spectral method for time domain model identification of non-stationary series," abbreviated the AR-SPECTRAL-TIME-ID approach.

Part III

8. A model of statistics for organization of graduate instruction

Statistics programs in American universities seem to have a history of faculty infighting over the question of what should be the organization of graduate instruction in statistics. Being an eternal optimist, I am hopeful that in the 1980's the academic statisticians may be able to reach a national consensus on the structure of their field for the purposes of advising students about the diversity of educational (and consequently career) paths they can follow.

Statistics programs have tended to think of graduate statistical education as being composed of two dimensions. In more theoretical programs, the dimensions were probability theory and statistical theory. In more applied programs, the dimensions were statistical theory and statistical methods. I propose now that we think in terms of four dimensions:

I Probability model oriented

II Statistical theory oriented

III Statistical methods oriented

IV Related theoretical and applied areas

These constitute four groups in which the various areas of statistics may be placed; a Master's or Ph.D. degree in Statistics would require the student to divide his coursework almost equally between the four groups.

To list areas of statistics, I adopt as a definition of an area that it is a field on which: (1) an academic program would offer a "how" (or applied or methods) course, and a "why" (or theory) course; and

(2) nationally-advertised short courses seem to be frequently offered.

A possible allocation of areas of statistics among four groups is as follows.

Group I. Probability Model Oriented.

1. Probability
2. Stochastic Processes
3. Time Series
4. Biometry, Biostatistics
5. Reliability
6. Operations Research, Mathematical Programming
7. Advanced Probability and Stochastic Processes

Group II. Statistical Theory Oriented.

1. Parametric statistical inference and estimation theories
2. Nonparametric
3. Multivariate
4. Contingency tables, discrete data analysis
5. Advanced statistical theory (including decision theory and sequential)

Group III. Statistical Methods Oriented.

1. Design
2. Regression, Linear Models
3. Sampling
4. Advanced Statistical Methods (including response surfaces and non-linear regression)

5. Statistical computing
6. Data analysis

Group IV. Related Theoretical and Applied Disciplines.

1. Introductory Statistics
2. Applicable mathematics (advanced calculus, matrix theory, real variables, complex variables, functional analysis)
3. Computer Science and Numerical Analysis
4. Econometrics and Sociometrics
5. Geostatistics
6. Signal Processing
7. Other subject matter areas applying statistics

Group V. Statistical Consulting and Collaboration.

Statistical consulting in the university has several roles, including (1) the delivery of statistical care (in the sense of helping researchers apply contemporary statistical computational methods), and (2) the formulation of statistical problems on which research is required to meet the needs of clients requiring statistical care. To heighten the awareness of researchers that statistics can matter, I often facetiously suggest that a statistical consulting center have the subtitle "statistical massage parlor: probability models introduced." More seriously, I suggest that they have the subtitle "statistical science: the partner, not the servant, of professional science."

REFERENCES

Box, G. E. P. (1976). Science and statistics. Journal of the American Statistical Assn., 71, 791-799.

Federer, W. T. (1978). Some remarks on statistical education. The American Statistician, 32, 117-121.

Healy, M. J. R. (1978). Is statistics a science? Journal of the Royal Statistical Society, Series A, 141, pp. 385-393.

Parzen, E. (1979). Nonparametric statistical data modeling. Journal of the American Statistical Assn., 74, March 1979.

Parzen, E. (1979). A density-quantile function perspective on robust estimation, R. Launer and G. Wilkinson, editors, Robust Estimation Workshop Proceedings, Academic Press: New York.

Parzen, E. (1979). Quantile function version of Bayes theorem.

Parzen, E. (1979). Density-quantile approach to the nonparametric two sample problem.

Parzen, E. (1979). Time series modelling, spectral analysis, and forecasting.

Tukey, E. (1977). Exploratory Data Analysis, Addison Wesley: Reading, Mass.

Tukey, J. W. (1962). The future of data analysis. Ann. Math. Statist. 33, 1-67.

Watts, D. G. (1968). The Future of Statistics, Academic Press: New York.